Learning SAS: Performing Univariate Analysis by Group Using PROC UNIVARIATE

Authored by

Mohammed looti

May 8, 2026

In the complex world of statistical data processing, deriving summary metrics not just for an entire dataset, but for distinct subgroups within it, is often essential for insightful analysis. The **PROC UNIVARIATE** procedure in **SAS** stands as a fundamental tool, designed to calculate a comprehensive array of **descriptive statistics** for numeric variables, providing the foundation for **univariate analysis**. However, its true power in comparative analysis is unlocked when it is combined with the **BY statement**, an indispensable command that instructs **SAS** to perform the analysis independently for every unique value found in a designated categorical grouping variable.

This methodology grants researchers and data analysts the capability to delve into the intrinsic characteristics of various data segments, enabling robust comparative analysis and the swift identification of patterns or anomalies specific to a subgroup. Regardless of whether your task involves scrutinizing performance metrics across disparate departments, comparing health biomarkers between clinical treatment cohorts, or analyzing financial returns segregated by geographical region, the synergistic application of **PROC UNIVARIATE** and the **BY statement** significantly streamlines the entire process. This combination ensures that you obtain detailed, segregated statistical summaries tailored precisely to each distinct category within your dataset, maximizing the clarity of your findings.

Mastering the Syntax for Grouped Univariate Analysis

The core structure required to execute **PROC UNIVARIATE** with group segmentation via the **BY statement** in **SAS** is both elegant and exceedingly powerful. The process fundamentally requires the analyst to specify the statistical procedure, identify the dataset intended for analysis, and, most crucially, nominate the grouping variable that dictates how the data will be partitioned into subsets for

independent processing. Once executed, the procedure generates a distinct and comprehensive set of univariate analysis results for every unique level or value present in the specified grouping variable.

The general syntax shown below is the standard structure for this powerful operation. It is vital to understand a core requirement: the
BY statement
mandates that the input dataset must be physically sorted according to the grouping variable before the procedure is invoked. This prerequisite exists because
SAS
processes data sequentially, requiring contiguous blocks of identical grouping variables. Therefore, while not mandatory in all cases, incorporating a
PROC SORT
step immediately prior to
PROC UNIVARIATE
is considered best practice and is often essential for correct execution, preventing potential errors or incomplete analysis.

```
proc univariate data=my_data normal;
by group_variable;
run;
```

In this illustrative syntax, the term `my_data` is the placeholder for the input dataset containing the raw information you intend to analyze, and `group_variable` represents the specific categorical field that will be used to segment your data. The inclusion of the optional
`NORMAL`
keyword is particularly useful, as it instructs
PROC UNIVARIATE
to execute formal tests for normality. This provides critical supplementary information regarding the statistical distribution of the numeric variables within each of the segregated groups, offering deeper insights beyond standard central tendency metrics.

Preparing the Dataset: An Illustrative Example in SAS

To effectively demonstrate the practical application and significant utility of coupling
PROC UNIVARIATE
with the
BY statement,
we will construct a detailed, hypothetical dataset focused on basketball player statistics. This data structure will allow us to clearly illustrate how to calculate comprehensive

descriptive statistics
for key player performance metrics, specifically categorized and grouped according to the players' respective teams. This segmentation is crucial for drawing meaningful comparisons.

The following
DATA step
code block is used to meticulously create our sample dataset, which we have named `my_data`. This dataset is designed to contain three core variables: `team`, which is a character variable identifying the player's affiliation; `points`, a numeric field tracking points scored; and `rebounds`, another numeric field quantifying rebounds achieved. The **`INPUT statement`** is utilized to precisely define the structure and data type for each of these variables, while the
`DATALINES statement`
provides a convenient mechanism to embed the actual raw data records directly within the
SAS
program for immediate execution.

```
/*create dataset*/
data my_data;
input team $ points rebounds;
datalines;
A 12 8
A 12 8
A 12 8
A 23 9
A 20 12
A 14 7
A 14 7
B 20 2
B 20 5
B 29 4
B 14 7
B 20 2
B 20 2
B 20 5
;
run;

/*view dataset*/
proc print data=my_data;
```

Following the successful creation and population of the data, a straightforward
PROC PRINT
statement is executed. The sole purpose of this initial step is to display the contents of the recently constructed `my_data` dataset in the output window. This immediate visualization and verification are paramount, serving as a critical checkpoint to confirm the data's integrity, ensuring that it is correctly structured, and verifying that it is fully prepared for the subsequent, rigorous statistical processing steps.

Obs	team	points	rebounds
1	A	12	8
2	A	12	8
3	A	12	8
4	A	23	9
5	A	20	12
6	A	14	7
7	A	14	7
8	B	20	2
9	B	20	5
10	B	29	4
11	B	14	7
12	B	20	2
13	B	20	2
14	B	20	5

Executing PROC UNIVARIATE with the BY Statement

With the example dataset now verified and ready, the next step involves calculating
descriptive statistics
for the numeric variables, namely `points` and `rebounds`, while ensuring the results are distinctly grouped by the `team` variable. This crucial segregation step will produce separate statistical summaries for Team A and Team B, which is the exact requirement for performing a direct and meaningful comparison of the performance metrics between the players of the two teams.

As previously emphasized, the correct functioning of the
BY statement
in conjunction with
PROC UNIVARIATE
absolutely requires the dataset to be sorted by the grouping variable--in this instance, `team`. While

the primary example code block focuses on the analysis step, the preceding sorting step is non-negotiable for production environments. A typical preliminary command would be `PROC SORT DATA=my_data OUT=my_data; BY team; RUN;`. Once the data is ordered, the
PROC UNIVARIATE
statement then instructs
SAS
to execute a detailed
univariate analysis
on every numeric variable present in the dataset, meticulously segregated based on the levels established by the `team` variable.

```
proc univariate data=my_data;
by team;
run;
```

Upon successful execution of this procedure,
SAS
will generate a series of distinct output sections. For each unique team identified in the `team` variable, the output will contain a complete
univariate analysis
for both the `points` and `rebounds` variables. This highly structured and segmented output significantly simplifies the task of comparing critical statistical measures--such as means, medians, standard deviations, and quartiles--across the different groups, thereby providing a clear and immediate picture of each subgroup's unique characteristics and performance profile.

Interpreting the Output from Grouped Univariate Analysis

The output resulting from the execution of
PROC UNIVARIATE
when paired with the
BY statement
is a detailed, multi-section report, rigorously segmented according to every level of the specified grouping variable. For our basketball player dataset, this analytical process yields a series of extensive
descriptive statistics
tables, where each table corresponds precisely to a specific team and a specific performance variable. The comprehensive nature of this output includes standard measures of central tendency (such as the mean, median, and mode), measures of dispersion (including standard deviation, variance, and range), and crucial indicators of distribution shape (like skewness, kurtosis, and

various quantiles).

Specifically, the procedure organizes and generates the following four key statistical summaries, illustrating the power of the grouping functionality:

Descriptive statistics for the **points** variable, analyzed exclusively for team **A**.
Descriptive statistics for the **rebounds** variable, analyzed exclusively for team **A**.
Descriptive statistics for the **points** variable, analyzed exclusively for team **B**.
Descriptive statistics for the **rebounds** variable, analyzed exclusively for team **B**.

Each of these segregated summaries offers an in-depth view into the numeric characteristics and distributional properties of the chosen variable within its respective group. For instance, an analyst can quickly identify the average points scored by players in Team A, the median rebounds achieved by Team B, and the degree of variability in performance within either team. This granular level of statistical detail is absolutely invaluable for accurately understanding within-group trends, confidently performing group-to-group comparisons, and making data-driven decisions based on specific subgroup profiles.

A visual snippet of the typical output is presented below, specifically highlighting the
descriptive statistics
calculated for the `points` variable for team
A.
This image clearly demonstrates the standard format and the rich type of statistical information that is routinely presented by
PROC UNIVARIATE.

The UNIVARIATE Procedure
Variable: points

team=A

Moments			
N	7	Sum Weights	7
Mean	15.2857143	Sum Observations	107
Std Deviation	4.42396073	Variance	19.5714286
Skewness	1.22128564	Kurtosis	-0.0509457
Uncorrected SS	1753	Corrected SS	117.428571
Coeff Variation	28.9417992	Std Error Mean	1.67209999

Basic Statistical Measures			
Location		Variability	
Mean	15.28571	Std Deviation	4.42396
Median	14.00000	Variance	19.57143
Mode	12.00000	Range	11.00000
		Interquartile Range	8.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	9.141627	Pr > |t|	<.0001
Sign	M	3.5	Pr >= |M|	0.0156
Signed Rank	S	14	Pr >= |S|	0.0156

The subsequent sections of the report, covering the other variables and teams, maintain this consistent structure, enabling systematic and efficient review and interpretation. The second image provides an additional example of the detailed statistical tables generated, which typically includes various calculated moments, fundamental statistical measures, and often includes the results of tests for normality. This holistic view ensures a deep understanding of the data's distribution within each and every defined group.

Quantiles (Definition 5)	
Level	Quantile
100% Max	23
99%	23
95%	23
90%	23
75% Q3	20
50% Median	14
25% Q1	12
10%	12
5%	12
1%	12
0% Min	12

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
12	3	12	3
12	2	14	6
12	1	14	7
14	7	20	5
14	6	23	4

Refining Your Analysis with the VAR Statement

While conducting
univariate analysis
on every numeric variable within each defined group provides a comprehensive overview, data analysts frequently encounter scenarios where their interest is strictly limited to the
descriptive statistics
of only one or a small, select group of specific variables. In these focused analytical situations, the
VAR statement
becomes an indispensable mechanism for precisely narrowing the operational scope of the analysis.

The
VAR statement
allows the user to explicitly list only those variables for which the calculation of
descriptive statistics

is required. This strategic use of the statement accomplishes several objectives: it significantly streamlines the output by eliminating extraneous or irrelevant information, thereby enhancing clarity, and it can substantially improve processing efficiency, particularly when dealing with extremely large datasets that contain a multitude of numeric variables unrelated to the current research question. By combining the restrictive power of the
VAR statement
with the segmentation capability of the
BY statement,
you gain highly precise and controlled execution over which variables are analyzed on a group-by-group basis.

For instance, if the analytical goal is solely to obtain the
descriptive statistics
for the `points` variable, categorized and grouped by `team`, the simplified and focused syntax required would be:

```
proc univariate data=my_data;
var points;
by team;
run;
```

This refined approach guarantees that the resulting output is concentrated entirely on addressing your specific analytical queries. The system offers considerable flexibility: you can specify multiple variables within the
VAR statement
as needed, and even define a sequence of grouping variables in the
BY statement,
provided that the dataset has been correctly sorted using the same hierarchy. This inherent adaptability is what renders
PROC UNIVARIATE
an exceptionally versatile and powerful tool for meeting diverse data analysis requirements within the
SAS
programming environment.

Summary and Path to Further Statistical Exploration

The seamless integration of
PROC UNIVARIATE
with the

BY statement
in
SAS
provides analysts with a potent and highly efficient methodology for conducting thorough
univariate analysis
across distinct, defined subgroups within a dataset. This capability is fundamentally important for conducting rigorous comparative studies, achieving a deep understanding of group-specific characteristics, and successfully isolating underlying statistical patterns that would otherwise remain obscured or diluted within an aggregated, overall analysis.

By judiciously employing the
BY statement
to accurately segment your data and utilizing the optional
VAR statement
to precisely select relevant variables, you can meticulously tailor your statistical investigations to directly address highly specific research questions. It is imperative to always recall and execute the necessary sorting of your data by the grouping variable(s) prior to implementing the
BY statement
to guarantee both accurate execution and statistically valid results.

We strongly encourage all users to actively experiment with various datasets, diverse variables, and the extensive range of options available within
PROC UNIVARIATE
to fully internalize and grasp its immense analytical potential. Achieving mastery of this essential procedure will profoundly enhance your capability to perform rigorous, detailed, and truly insightful data analysis within the
SAS
environment.

The following tutorials explain how to perform other common tasks in
SAS: